40266

National Library of Canada | Bibliothèque nationale du Canada

CANADIAN THESES ON MICROFICHE | THÈSES CANADIENNES SUR MICROFICHE

NAME OF AUTHOR/*NOM DE L'AUTEUR* Chen-Tung Ou

TITLE OF THESIS/*TITRE DE LA THÈSE* On the distribution of query-document correlations

UNIVERSITY/*UNIVERSITÉ* University of Alberta

DEGREE FOR WHICH THESIS WAS PRESENTED/ *GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE* M. Sc

YEAR THIS DEGREE CONFERRED/*ANNÉE D'OBTENTION DE CE GRADE* 1978

NAME OF SUPERVISOR/*NOM DU DIRECTEUR DE THÈSE* Clement T. Yu

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

DATED/*DATÉ* Sept 22 '78 SIGNED/*SIGNÉ* Chen-Tung Ou

PERMANENT ADDRESS/*RÉSIDENCE FIXE* #88, Lane 79
Chung-shan Rd.
Taipan, Taiwan 700

NL-91 (3-74)

National Library of Canada

Cataloguing Branch
Canadian Theses Division

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada

Direction du catalogage
Division des thèses canadiennes

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

THE UNIVERSITY OF ALBERTA

ON THE DISTRIBUTION OF QUERY-DOCUMENT CORRELATIONS

by

© CHEN-TUNG OU

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

IN

COMPUTING SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

FALL, 1978

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled ON THE DISTRIBUTION OF QUERY-DOCUMENT CORRELATIONS

submitted by CHEN-TUNG OU

in partial fulfilment of the requirements for the degree of Master of Science in Computing Science.

Supervisor

Date August 29, 1978

TO MY PARENTS

ABSTRACT

In documents retrieval system, it is a common practice to organize the documents into clusters and search only a few promising clusters. In such an environment it is of interest to estimate, with respect to a given query, the number of "desired documents" in each cluster. One of the reasonable approaches to this problem is to predict the distribution of query-document correlations by using the cluster representative. The purpose of this thesis is to evaluate the accuracy of a number of theoretical distributions in predicting the distribution of query-document correlations. They are the C-function, the Poisson distribution, and the normal distribution. Two modified C-functions are also proposed to alleviate some of the difficulties inherent in the use of the C-function. Since terms are found to be dependent on each other, a transformation that is expected to remove the dependencies between terms is attempted. The distribution of correlations for the newly transformed data is also derived.

ACKNOWLEGEMENT

I wish to express my sincere appreciation to my supervisor, Dr. C. T. Yu, for his patience, guidance, and financial support throughout the preparation of this thesis. His many readings of earlier drafts and his suggestions were most helpful. Thanks are also due to the other members of the examination Committee, Dr. I-Ngo Chen, and Dr. C. Davis of Library Science for the time they spend reading this thesis and their comments.

I am also grateful to Vijay Ragahvan for his many helpful suggestions and our fruitful discussions about the content and the writing style of this thesis.

Financial support was also provided by the Department of Computing Science, in the form of Teaching Assitanship.

TABLE OF CONTENTS

page

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 THE DISTRIBUTION OF QUERY-DOCUMENT CORRELATIONS ON BINARY DATA 4

2.1 Some Characterizations of the C-Function 5

2.1.1 The C-Function and Its Probability Generating Function 6

2.1.2 The Mean Value and the Variance of the C-Function 7

2.1.3 Approximations of the C-Function by other Probability Functions 9

2.2 Experimental Results 12

2.2.1 Evaluation of the Accuracy in Predicting the Actual Distribution by the C-Function 13

2.2.2 The C-Function with Modifications 16

2.2.2.1 The Modified C-Function with Dependent Terms Being Combined 16

2.2.2.2 The Modified C-Function with a Classified Collection 18

2.2.3 Evaluation of the Accuracy in Predicting the Actual Distribution by the Poisson Distribution and by the Normal Distribution 21

2.2.3.1 Prediction by the Poisson Distribution 21

- cont'd -

page

2.2.3.2 Prediction by the Normal Distribution 23

CHAPTER 3 THE DATA TRANSFORMATION AND THE DISTRIBUTION OF QUERY-DOCUMENT CORRELATIONS ON WEIGHTED DATA 24

3.1 The Data Transformation 24

3.1.1 The Concept of the Transformation ... 25

3.1.2 The Procedure of the Transformation . 27

3.1.3 The Characterization of the Transformation 29

3.1.3.1 The Number of Terms after Transformation 29

3.1.3.2 Other Remarks on the Transformation 30

3.2 The Distribution of Correlations on Weighted Data 31

3.2.1 The Actual Distribution of correlations 32

3.2.2 The Theoretical Distribution - a Generalized C-Function For Weighted Data 33

3.3 Experimental Result 35

CHAPTER 4 CONCLUSION 36

- cont'd -

page

TABLE 1 39

TABLE 2 40

TABLE 3 41

BIBLIOGRAPHY 42

CHAPTER 1

INTRODUCTION

An information retrieval system consists of a set of documents in machine-readable form stored to provide information services to various users. A document or a query can be represented by an n-dimensional vector whose ith component indicates the importance of the ith term in the document or the query.

In response to a given query, the system may compute the closeness or the correlation between each document and the query in terms of a chosen similarity function (for example, a simple matching function computes the number of terms in common between the document and the query.) Documents which have correlation greater than a given threshold are then retrieved. We refer to this set of documents as the "desired documents". Since the number of documents in the system is usually very large (for example, a stored collection of books and journals in a library), it is impractical to examine each document and determine its closeness to the query. One way to remedy this situation is to divide the data base into clusters and examine only the clusters which have high probabilities of containing many desired documents. A cluster consists of "similar" documents and a representative can be constructed for each cluster to provide the required probability (see section 2.1.) Besides

being useful in the selection of promising clusters, an estimate of the number of desired documents in each cluster can also be used for some other applications in information retrieval such as the following:

(1). In order to reduce the amount of computing time, the system may search only few related clusters. In such a situation, the user may be interested in knowing the percentage loss of desired documents. The estimation of desired document will allow the system to provide this information. If this percentage is too high, the user may require the system to search more clusters.

(2). This estimate will also enable the system to give the user information about the number of documents that are likely to be retrieved. If this number is too large, the user may decide to add more constraints in his query, thereby reducing the searching time as well as the number of documents to be retrieved.

With respect to a given query, a distribution of correlations of a cluster of documents represents the (relative) frequencies of documents at different correlation values. If the distribution of correlations is known and a threshold of retrieval is given, then the number of documents which have correlation greater than the threshold can easily be obtained. Therefore, it is desirable to obtain the distribution of correlations in order to estimate the desired documents in each cluster. Several theoretical

distributions, namely, the C-function [2], the Poisson distribution [9,10] and the normal distribution [11], have been used to describe the distribution of correlations. The objective of this thesis is to evaluate the accuracy of these theoretical distributions and modify the C-function to alleviate some of the difficulties in the use of the C-function.

In Chapter 2, some characterizations of the C-function are derived and its relationships to the Poisson and the normal distributions are discussed. Experiments are performed comparing these theoretical distributions to the actual distribution of correlations. Experimental results demonstrate that if the collection is large the C-function fits rather poorly to the actual distribution. This inaccuracy is largely due to the dependencies between terms (see section 2.1.1, for explanation of dependence.) Two modified C-functions that take dependencies of terms into consideration are proposed. The accuracy of the theoretical distribution is substantially improved.

In chapter 3, the principal component analysis [5] technique is attempted to remove the dependencies between terms. A theoretical distribution based on the C-function is developed for this newly transformed set of data (documents and queries.) Similar experiments of fitting the theoretical distribution to the actual distribution are subsequently performed. Substantial improvement in accuracy of the estimation of the distribution is achieved.

CHAPTER 2

THE DISTRIBUTION OF QUERY-DOCUMENT CORRELATIONS ON BINARY DATA

It may be desirable to estimate the number of desired documents in a large collection. The terms "cluster", "class", "collection" will be used interchangable throughout this thesis. Consider a collection of N documents represented by the document matrix $M=(d_{ij})$ of n rows and N columns, where the ith row represents the ith term, the jth column represents the jth document D_j and $d_{ij}=1$ if document D_j has the ith term and $d_{ij}=0$ otherwise. Similarly a query Q is represented by a vector $Q=(q_1,\ldots,q_n)$ where $q_i=1$ if the query has the ith term and $q_i=0$ otherwise.

A measurment of the closeness between document D_j and the query Q is given by the simple matching function f:

$$f(D_j,Q)=\sum_{i=1}^{n} d_{ij}\cdot q_i.$$

The value of the function represents the number of terms in common between the document D_j and the query Q. This measure of closeness or similarity is referred to as the correlation of document D_j with query Q.

Once the correlations of all the documents with a given query are calculated, the number of documents having a correlation equal to i, for $i=1,\ldots,k$ (k is the number of nonzero terms in the query), are obtained. Thus, the

distribution of correlations with respect to that particular query can be exhibited in a histogram. The correlation from 0 to k are represented on the X-axis while the relative frequencies are plotted on the Y-axis. This process yields the actual distribution of correlations between the documents and the given query. However, it is not economical to compute the correlations between the documents in the cluster and the query. Hence, it is desirable to estimate this distribution by using the representative of the cluster without computing the correlations of documents.

In this chapter, many different theoretical distributions are used to predict this distribution of correlations. All these theoretical distributions are generated by using only the representative of a cluster. Experiments are subsequently performed on a number of data collections (selected from the SMART retrieval system [3]) to demonstrate how closely these predicting distributions fit the actual distribution.

2.1 Some Characterizations of the C-function

A function called the C-function was proposed [2] to predict the distribution of correlations of documents with respect to a given query. The details of the C-function and some related properties are discussed in this section.

2.1.1 The C-function and its Probability Generating Function:

Consider a class of N documents and a query to be represented by a document matrix M and a query vector Q as described earlier. The occurrence or the non-occurrence of the ith term in the documents can be represented by $\underline{X}_i = (d_{i1},\ldots,d_{iN})$ where d_{ij} is 1 if the ith term occurs in the jth document and 0 otherwise. Associated with $\underline{X}_i$, we define a Bernoulli random variable X_i' with parameter $p_i' = \sum_{j=1}^{N} d_{ij}/N$. This is the probability that a document in the class has the jth term. Thus a set of n random variables $\{X_1',\ldots,X_n'\}$ is obtained in which X_i' describes the occurrence of the ith term in the class of documents. The representative of a cluster is defined as $R=(p_1',p_2',\ldots,p_n')$.

Let Q be the given query containing k non-zero terms. Let $X_1,\ldots,X_k$ be the k independent random variables corresponding to the non-zero terms of query Q with X_i taking on value 1 with probability p_i and value 0 with probability $1-p_i$. Since random variables correspond to the terms of query, the terms "random variable" and "term" have the same meaning throughout this thesis. The independencies between terms imply that the occurrence of a term in a document does not affect the occurrence of the other terms. Let $X = \sum_{i=1}^{k} X_i$, then the probability that X has the value i is the relative frequency of documents whose correlation equals i and is designated by

$$C(p_1,\ldots,p_k;i)$$

$$= \sum_{\binom{k}{i}} \left(\prod_{h=1}^{i} p_{g(h)}\right)\cdot\left(\prod_{h=i+1}^{k} (1-p_{g(h)})\right) \quad \text{for } i=1,\ldots,k$$

$$= 0 \qquad \text{otherwise}$$

where g is a permutation of the integers {1,2,...,k} and the summation is taken over all $\binom{k}{i}$ combinations [2].

The C-function can also be described by means of its probability generating function. The probability generating function of each X_i is given by a polynomial $p_i s+(1-p_i)$. Since the generating function of the sum of independent random variables equals the product of the generating functions of these random variables [4], the probability generating function of $X = \sum_{i=1}^{k} X_i$ is $g(s) = \prod_{i=1}^{k} (p_i s+(1-p_i))$ where the coefficient of s gives the relative frequency of documents at correlation = i. This generating function will be used to derive the mean and the variance of the C-function. To compute the C-function efficiently, we shall simply multiply the probability generating functions of all the terms.

2.1.2 The Mean Value and the Variance of the C-function

It is of interest to derive the mean value and the variance of the distribution of correlations of a cluster of documents with respect to a given query. This mean value or the average query-document correlation may be viewed as a measurement of closeness between the "average" document in the cluster and the query. Clusters with larger average

query-document correlations are more related to the query. The variance indicates the average deviation of query-document correlations from the mean correlation. A small variance implies that most of the documents in the cluster have query-document correlations close to the mean. Clusters with large mean value and small variance of document-query correlations are, of course, strongly related to the query and contain many desired documents. These clusters should contain most of the documents of interest to the user.

If $g(s)$ denotes the probability generating function of integral valued random variable X, then the mean $E(X)=g'(1)$ and the variance $V(X)=g''(1)-g'(1)+(g'(1))^2$ where $g'(1)$ and $g''(1)$ are the first and second derivatives of g at $s=1$ [4].

The mean and the variance of the C-function can now be easily derived :

The mean is

$$E(X)= E\left(\sum_{i=1}^{k} X_i\right)= \sum_{i=1}^{k} (E(X_i))$$

$$= \sum_{i=1}^{k} g_i'(1)= \sum_{i=1}^{k} p_i ,$$

and the variance is obtained as

$$V(x)=V\left(\sum_{i=1}^{k} X_i\right)$$

$$= \sum_{i=1}^{k} V(X_i) \quad \text{(by independence of } X_i\text{'s)}$$

$$= \sum_{i=1}^{k} (g_i''(1)-g_i'(1)+(g_i'(1))^2)$$

$$= \sum_{i=1}^{k} p_i \cdot (1-p_i).$$

The variance of the C-function may be used to estimate the variance of the actual distribution of the query-document correlations. The following proposition shows that the mean of the C-function is equal to the mean of the actual distribution.

Proposition 2.1

Let a_i, i=0,...,k be the number of documents having i terms in common with the query and N be the number of documents in the cluster, the mean of the query-document correlations which is $\sum_{i=0}^{k} a_i \cdot i/N$ is equal to $\sum_{i=1}^{k} p_i$.

Proof :

Consider a matrix where the columns represent the k terms and the rows represent the N documents. The (i,j)th entry in this matrix is 1 if the ith document has the jth term and 0 otherwise. Then $\sum_{i=0}^{k} i \cdot a_i$ is the number of 1's in the matrix, counting them rowwise. If the 1's are counted columnwise, then we obtain $\sum_{i=1}^{k} N \cdot p_i$. Hence, the result follows. □

2.1.3 Approximations of the C-function by Other Probability Functions.

In addition to the C-function, the Poisson distribution [9,10,14,19] and the normal distribution [11] have been proposed by others for describing the distribution of correlations. The relationships of these distributions to

the C-function are discussed next.

The following proposition shows that the C-function approaches the Poisson distribution with the (same) mean, $\sum_{i=1}^{k} p_i$, whenever all p_i's are sufficiently small.

Proposition 2.2

$$\sum_{i=0}^{k} \left| c(p_1,\ldots,p_k;i) - e^{-\mu}\cdot\frac{\mu^i}{i!} \right| \le 2\sum_{i=1}^{k} p_i^2 \quad \text{where } \mu = \sum_{i=1}^{k} p_i .$$

Proof :

We first show that $|d_1(h) - d_2(h)| \le 2p_i^2$ (2.1)

where d_j, $1 \le j \le 2$, are respectively the density functions for a Bernoulli and a Poisson random variables on the ith term, both with parameter p_i, i.e.

$$d_1(h) = \begin{cases} 1 - p_i & \text{if } h=0 \\ p_i & \text{if } h=1 \\ 0 & \text{if } h=2,3,\ldots \end{cases} \qquad \text{and } d_2(h) = e^{-p_i}\cdot\frac{p_i^h}{h!}$$

The L.H.S. of (2.1)

$$= |(1-p_i) - e^{-p_i}| + |p_i - p_i e^{-p_i}| + \left|0 - e^{-p_i}\cdot\frac{p_i^2}{2!}\right| + \ldots$$

$$= [e^{-p_i} - (1-p_i)] + [p_i(1-e^{-p_i})] + \left[e^{-p_i}\frac{p_i^2}{2!}\right] + \ldots$$

(since $1 \ge e^{-x} \ge 1-x$, for any $x \ge 0$)

$$= e^{-p_i} - 1 + p_i + p_i - p_i e^{-p_i} + e^{-p_i}(e^{p_i} - 1 - p_i) + \ldots$$

(since $e^{p_i} = 1 + p_i + p_i^2/2! + \ldots$)

$= 2p_i(1-e^{-p_i}) \leq 2p_i$ (since $e^{-x} \geq 1-x$, for any $x \geq 0$).

The proposition is proved by applying (2.1) to exercises (12) and (14) on p. 286 of vol. 2 in [4]. □

The next two propositions demonstrate that the C-function can be approximated by the normal distribution whose mean is $\sum_{i=1}^{k} p_i$ and whose variance is $\sum_{i=1}^{k} p_i(1-p_i)$ when all the p_i's are near 1/2.

Proposition 2.3

$$\left| \sum_{i=k}^{\infty} C(p_1,\ldots,p_k;i) - \int_{\frac{h-k\mu}{\sigma}}^{\infty} 1/\sqrt{2\pi}\cdot e^{-z^2/2}dz \right| < 6A(p_1,\ldots,p_k)$$

where $A(p_1,\ldots,p_k) = \max_{1\leq i\leq k} \{p_i^2+(1-p_i)^2\}/\sigma$

and $\sigma^2 = \sum_{i=1}^{k} p_i(1-p_i)$, $k\mu = \sum_{i=1}^{k} p_i$.

Proof:

By [4, p. 544, theorem 2], we have

$$\left| \sum_{i=k}^{\infty} C(p_1,\ldots,p_k;i) - \int_{\frac{h-k\mu}{\sigma}}^{\infty} 1/\sqrt{2\pi}\, e^{-z^2/2}dz \right|$$

$$\leq 6\frac{\sum_{i=1}^{k} p_i(1-p_i)\cdot(p_i^2+(1-p_i)^2)}{\sigma^3}$$

$$\leq 6 \max_{1\leq i\leq k} \{p_i^2+(1-p_i)^2\}/\sigma.$$

Proposition 2.4

let A be the function of $p_1,\ldots,p_k$ define in proposition 2.3. Then $A(1/2,1/2,\ldots,1/2) \leq A(p_1,p_2,\ldots,p_k)$ for $0 \leq p_i \leq 1$; $1 \leq i \leq k$.

Proof :

It is easy to see that the numerator of A does not increase and the denominator of A strictly increases as p_i increase in $[0,1/2)$. Since $A(p_1,\ldots,p_k) = A(1-p_1,\ldots,1-p_k)$, the desired result follows. □

2.2 Experimental Results

As mentioned in the last section, different theoretical distributions, namely, the C-function, the Poisson distribution and the normal distribution are used to predict the actual distribution of correlations of the documents with respect to a given query. Experiments are performed to find out the accuracy of such predictions. The Chi-square test of the goodness of fit is used as a criterion to judge the accuracy of a given theoretical distribution to the actual destribution. The null hypothesis is that the actual distribution follows the theoretical distribution. Initially, there are (k+1) classes or cells corresponding to each of the correlation values from 0 to k. By the theory of Chi-square test, the neighbouring classes are combined until the smallest expected frequency in each class is at least 5. Experimental results are based on a 5% level of significance using m-1 degrees of freedom, where m is the number of classes that result after neighboring classes are combined.

A "data collection" in the SMART system [3] consists of a class or a collection of documents and a set of queries. With respect to each query, a Chi-square test is performed to fit the theoretical distribution to the actual distribution. Although the accuracy may be judged by the Chi-square quantity obtained, our discussion is based on the percentage of the good fits of tests(the number of tests = the number of queries in the collection.) The following collections are chosen from the SMART system [3]. They are ABIABTH, ADINUL, CRN2NUL, CRN2TH, CRN4NUL, MEDNUL, and CRN14TH.

2.2.1 Evaluation of the Accuracy in Predicting the Actual Distribution by the C-function.

The first theoretical distribution to be tested is the C-function. The results are given in the first column of Table 1.

It is found (Table 1) that the actual distribution and the C-function distribution are close to each other for most queries found in small collections, as in collection ADIABTH (97.1% good fits) and ADINUL (88.6% good fits) [3]. They differ substantially for most queries in the larger collections. For example, only 4.9% of the queries in the actual distributions follow the C-function in collection CRN14TH [3].

Two reasons are suspected for this bad result:

(1). Terms are not independent.

It shoud be pointed out that the C-function is derived based on the assumption that all terms are independent. If this condition holds, the C-function would be a good approximation to the actual distribution. However, this ideal condition may not be satisfied by the actual data. The contingency table test based on the Chi-square test is used to find out the degree of independence of the terms of the queries. Experiments are performed for some of the queries in the CRN14TH collection. Sample results tested on the CRN14TH collection are given in Table 2. 249 (34%) out of a total of 734 term pairs (a sample) are found to be dependent in the CRN14TH collection. In contrast, 69 (7%) out of 988 term pairs (a sample) are dependent in the ADINUL collection.

Due to the dependence between terms found in large collections, the C-function is rather different from the actual distribution. The reason that terms in large collections are likely to be dependent is discussed in section 2.2.2.2.

(2). The limitation of the Chi-square test.

Since the Chi-square distribution has been established as the limiting distribution as the size of a sample increases [14], it is clear that a large sample is recommended for the test. If the actual distribution follows exactly a certain theoretical

distribution, then the more document we sample, the better will be the fit of the actual distribution to the theoretical distribution. On the other hand, if the actual distribution does not follow a certain theoretical distribution exactly, but just close to it, then the actual distribution from a small sample will tend to fit the theoretical one easily while the actual one from a large sample is rather difficult to fit the theoretical distribution.

In his original paper, Karl Pearson [18] was aware of this. He writes, "Nor again does it appear to follow that if the number[1] be largely increased the same curve will still be a good fit. Roughly, the Chi-square's[2] of two samples appear to vary for the same grouping as their total contents. Hence, if a curve be a good fit for a large sample, it will be good for a small one, but the converse is not true, and a larger sample may show that our theoretical frequency gives only an approximate law for samples of a certain size."

With the above concept in mind, it is unfair to compare results of different size of collection. Therefore, different sizes of samples were then randomly selected from the CRN14TH collection. Similar experiments were performed on these samples. Results given in Table 3 demonstrate that the C-function approximates the actual distribution much better for

[1] the size of sample.
[2] the Chi-square quantity.

small samples. The results of poor approximation in large collection is partly due to this limitation of the Chi-square test. This difficulty with the Chi-square test is encountered throughout all experiments in this thesis.

It should also be pointed out that the C-function generally under-estimates the "beginning part" and the "ending part" of the actual distribution. The beginning part of the distribution indicates the frequencies at small correlation values while the ending part refers to those having large correlations.

2.2.2 The C-function with Modifications

In view of the fact that some terms are dependent in large collections, the C-function is modified to take into consideration this dependence of terms.

2.2.2.1 The Modified C-function with Dependent Terms Being Combined.

The C-function is modified as follows: The dependence of a pair of terms is measured by the Chi-square quantity obtained from contingency table test. This quantity is treated as a measurement of the degree of dependence of the two terms. The term pair with a large quantity implies a high dependence between the two terms. As an approximation, terms are assumed to depend on at most one other term. In the experiments to be described, a term is combined with the

most dependent term which has not yet been combined so far. For each term pair, a generating function is formed. Since terms in different term pairs are assumed to be independent, the generating functions of different term pairs and the generating functions of each independent term (if any) are multiplied together to obtain the generating function for the set of all terms.

Experimentally, this modified C-function is produced as follows:

Consider a set of terms $\{X_1,\ldots,X_k\}$ in which some are dependent while the remainder are independent. The two most dependent terms, X_i and X_j, are first extracted from the set of dependent terms. Let $\{a_{ij},b_{ij},c_{ij}\}$ be the actual frequencies of documents having both X_i and X_j, either X_i or X_j, and neither X_i nor X_j respectively. The polynomial $a_{ij}s^2+b_{ij}s+c_{ij}$ is then used as the probability generating function of the combination of the terms X_i and X_j.

After X_i and X_j are removed the above process is repeated on the remaining terms until there are no dependent term-pairs left. For each dependent term-pair, a generating function is obtained. We next compute the generating functions for each of the remaining terms. By multiplying all the generating functions, the desired expected distribution is obtained. This is an approximation to the actual distribution with pairwise dependence of terms taken into consideration. The experimental results of approximating the actual distribution by the above expected

distribution are given in the second column of Table 1. The results (for example, improved from 4.9% to 24.4% good fits for the CRN14TH collection) demonstrate a significant improvement over those obtained when the C-function was used.

The difference between this modified distribution and the C-function is that, instead of using the product of the generating functions of X_i and X_j, the function $a_{ij}s^2+b_{ij}s+c_{ij}$ is used. It is clear that a more complicated cluster representative is required for this modification which provides the probabilities that two terms co-occur and exactly one term occurs. This modification is performed for each pair of dependent terms identified. The error in the estimation of the distribution for X_i+X_j due to the dependence between terms such as X_i and X_j is eliminated. Thus, the modified distribution is a better approximation of the actual distribution than the C-function. Note, however, that the dependency of the terms in each term pair is eliminated, but the dependencies between terms in different term pairs still exist.

Since this modification takes into account the dependencies between two terms, the better prediction obtained suggests that the terms are not independent in some data collections being tested. To get a better prediction a more complicated model is required which takes into account dependencies involving more than two terms. Another alternative is to remove the dependencies between terms

which will be discussed in Chapter 3.

2.2.2.2 The Modified C-function with a Classified Collection

If the terms in a large collection are likely to be dependent whereas terms can be expected to be independent in small collection of similar documents, then if a large collection can be split into small clusters of similar documents, the C-function would be a good prediction of the actual distribution on each small cluster.

In the following experiments, the theoretical distributions are produced as follows:
First, the whole document collection is divided into two sets, the relevant and the irrelevant documents, with respect to each query and the corresponding C-functions are obtained. Finally, these two distributions are combined as follows:

> the expected number of documents having i terms in common with a query = the expected number of relevant documents having i terms in common with the query + the expected number of irrelevant documents having i terms in common with the query, for i=0,...,k.

Experimental results of fitting the above distribution to the actual distribution are given in the third column of Table 1. A substantial improvement is achieved for the collections of median size, for example, CRN4NUL (improved from 17.4% to 45.2%) and MEDNUL (from 40% to 63.3%.) However, since the number of relevant documents is too small

(less than 1%) compared to the number of total documents in CRN14TH (1400 documents), the modification has little effect on the expected distribution. The improvement is, hence, not significant for this large collection.

It should be pointed out that the above results made use of the assumption that the terms in the relevant set are independent and those in the irrelevant set are also independent. The better prediction obtained implies that the terms in a collection of similar documents are more likely to be independent. We may assume that a class of documents with similar characteristics have independent terms. If a collection is large, documents in the collection are unlikely to be similar. Hence, the terms in large collections are likely to be more or less dependent. If a large collection can be classified properly into some small clusters of homogeneous documents, this modified C-function is likely to be a good approximation to the actual distribution.

Since the information available on the data collection is not sufficient to obtain clusters of homogeneous documents, the whole collection is split only into two sets (the relevant and the irrelevant documents.) However, it is possible to apply clustering algorithms or employ factor analysis [12,13] to partition a given set of documents.

2.2.3 Evaluation of the Accuracy in Predicting the Actual Distribution by the Poisson Distribution and the Normal Distribution.

Damerau [9] has assumed that the occurrence frequency of a term in a large collection of documents follows the Poisson distribution. However, Swets [11] has described the distribution of correlations by a normal distribution in his study of the measure of effectiveness of an information retrieval system.

In this section, the accuracy of these two probability functions in predicting the actual distribution of correlations is considered.

2.2.3.1 Prediction by the Poisson Distribution.

In the following experiments, the Poisson distribution with mean, $\sum_{i=1}^{k} p_i$ is used to approximate the actual distribution. It is easy to show that the variance of the Poisson distribution $\sum_{i=1}^{k} p_i$, is larger than that of the C-function. It can also be shown that the relative frequency for the Poisson distribution is greater than that of the C-function at correlation = 0. The frequency is also greater at k if k, the number of the terms of the given query, is not too small (greater than 3) and the mean of p_i's is not too large (no larger than 1/2.)

Experimental results in column 4 of Table 1 show that the Poisson distribution is a better approximation than the

C-function for large collections (for example, 34.2% good fits compare to 4.9% by the C-function in CRN14TH.) As pointed out in section 2.2.1, the experiments show that the C-function usually under-estimates the beginning part and the ending part of the distribution of correlations. Since the Poisson distribution has its beginning and ending part greater than those of the C-function, a better approximation to the actual distribution is achieved.

The Poisson distribution is usually used to characterize a distribution of a random variable which can take on many values (for instance, the number of occurrences) but with small probability [17]. Since a large collection is likely to be more heterogeneous than a small collection and a particular term may not appear in too many documents, we assume that the probability that a term occurs in a document is small for a large collection. This satisfies the property of a Poisson distribution. Hence, Damerau's assumption that occurrence frequencies of a term follow a Poisson distribution should be considered meaningful only for a large collection of documents. Furthermore, the distribution of correlations follows the Poisson distribution by the addition property of the Poisson distributions of each term. In summary, a Poisson distribution is suitable for a large collection. Harter [10] in his recent study found that the single Poisson distribution is not adequate. The two-Poisson model and the linked two-Poisson model were recently proposed [19,10,14].

2.2.3.2 Prediction by the Normal Distribution.

In the following experiments a normal distribution with mean $\sum_{i=1}^{k} p_i$ and variance $\sum_{i=1}^{k} p_i(1-p_i)$ is used to approximate the actual distribution of correlations. Since the normal distribution is continuous, the area in the interval (i-1/2, i+1/2) is assumed to be the relative frequency of documents having correlation equal to i.

Experimental results given in column 5 of Table 1 show that the normal distribution does not fit the actual distribution, (0% in the CRN14TH collection) except for the ADINUL and ADIABTH collections. Experimental results show that the normal distribution has a worse approximation than the C-function. Since the p_i's are fairly small in most of the collections, except in the ADINUL collection, most of the documents have small correlations with the query. Hence, the distribution of correlations is rather skewed towards the ending part. A normal distribution which is symmetric, is thus not likely to be a reasonable fit to an actual distribution that is skewed.

Consequently, with small p_i's, Swets' [11] assumption of the normal distribution of correlations is questionable. In general, the p_i's tend to be small for a large collection, and as mentioned, Swets' [11] model is not suitable for such a collection. However, if the means of the p_i's are neither too large nor too small (the means of the p_i's are usually greater than 0.25 in the ADINUL collection) the normal distribution may be appropriate.

CHAPTER 3

THE DATA TRANSFORMATION AND THE DISTRIBUTION OF QUERY-DOCUMENT CORREIATIONS ON WEIGHTED DATA

The experiments in Chapter 2 demonstrate that the dependencies between terms may cause the C-function to be rather different from the actual distribution of correlations. However, if the dependencies between terms can be removed, the C-function may fit the actual distribution well. We shall employ the principal component analysis [5], to transform a set of dependent terms to a set of uncorrelated terms. Although the uncorrelation between terms does not imply their independence, the two concepts may be assumed to be equivalent for many practical purposes. Similar experiments will be subsequently performed on a data set obtained after performing the transformation.

3.1 The Data Transformation

The transformation induced on the data by the method referred to above is described in this section. It transforms a set of dependent terms $\{X_1,\ldots,X_n\}$ to a new set of uncorrelated concepts $\{Y_1,\ldots,Y_n\}$. Each transformed term (new concept) Y_i corresponds to a linear combination of $\{X_1,\ldots,X_n\}$.

3.1.1 The Concept of the Transformation

The transformation is based on some fundamental theorems in linear algebra. Some preliminary ideas are introduced first.

(1) Every symmetric matrix S of rank n has n independent eigenvectors associated with its n eigenvalues [6].

(2) If these n independent eigenvectors are chosen to be of unit length, then the matrix $A=(a_{ij})$ formed with these n independent eigenvectors as the rows of A is orthogonal, that is, $A^T=A^{-1}$ and A can be used to transform this symmetric matrix S into a diagonal matrix as shown below

$$A \bullet S \bullet A^{-1} = D$$

where D is the diagonal matrix having the eigenvalues of S along the diagonal [7].

Consider a document matrix $M=(d_{ij})_{nN}$ where a row represents a term and a column represents a document. Define n random variables $\{X_1,\ldots,X_n\}$, associated with term vectors of M. Without loss of generality, we may assume that all X_i's are standardized, with zero mean and unit variance.

Let $S=(r_{ij})$ be the matrix of correlation coefficient where $r_{ij}=E[(X_i-\mu_i)(X_j-\mu_j)]/\sigma_i \cdot \sigma_j$, the correlation coefficient of X_i and X_j. μ_i, σ_i, and μ_j, σ_j are the mean and the standard deviation of X_i and X_j respectively. Since the mean of each X_i is 0 and variance is 1, r_{ij} is $E[(X_i \bullet X_j)]$.

Suppose that S_{nn} is of rank n (see 3.1.3.1, if the rank

of $S < n$.) By the above theorems in linear algebra, there exists an orthogonal matrix $A = (a_{ij})$ such that $A \cdot S \cdot A^{-1} = D$ where the ith row of A is the ith eigenvector of S and D is a diagonal matrix whose (i,i)th entry is the ith eigenvalue of S.

Define a set of n random variables $\{Y_1, \ldots, Y_n\}$ as n linear combinations of $\{X_1, \ldots, X_n\}$ as follows:

$$Y_i = \sum_{j=1}^{n} a_{ij} X_j \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3).$$

We now prove that these n new variables are uncorrelated and the variance of Y equals the ith eigenvalue of S.

Proposition 3.1

$r(Y_i, Y_j) = 0$ for $i,j = 1, \ldots, n$ $i \neq j$

$V(Y_i) = \lambda_i$ for $i = 1, \ldots, n$

where Y_i, Y_j are defined in (3), $r(Y_i, Y_j)$ is the correlation coefficient of Y_i and Y_j, $V(Y_i)$ is the variance, and λ_i is the ith eigenvalue of S.

Proof :

By othorgonality of A, $A^T = A^{-1}$

the (i,j)th entry of $A \cdot S \cdot A^{-1}$

$$= \sum_{p=1}^{n} (\text{ (i,p)th entry of } A \cdot S \text{ }) \cdot (\text{ (p,j)th entry of } A^{-1} \text{ })$$

$$= \sum_{q=1}^{n} \left(\sum_{p=1}^{n} a_{iq} \cdot r_{qp} \right) \cdot a_{jp}$$

$$= \sum_{p,q=1}^{n} a_{iq} \cdot r_{qp} \cdot a_{jp}$$

$\mathrm{Cov}(Y_i, Y_j)$ where $\mathrm{cov}(Y_i, Y_j)$ is covariance of Y_i, Y_j

$$= \mathrm{cov}\left(\sum_{q=1}^{n} a_{iq} \cdot X_q, \sum_{p=1}^{n} a_{jp} \cdot X_p \right)$$

$$= \sum_{p,q=1}^{n} a_{iq} \cdot a_{jp} \cdot r_{qp}$$ (since $\sigma_{X_p} = \sigma_{X_q} = 1$ where σ_{X_p} and σ_{X_q} are standard deviations of X_p and X_q.

The result follows.

3.1.2 The Procedure of Transformation.

The transformation based on the above concept can be carried out using the following steps:

<u>step 1:</u> Construct the correlation coefficient matrix.

$S = (r_{ij})_{nn}$ of $\{X_1, \ldots, X_n\}$ where r_{ij} is the correlation coefficient of X_i and X_j and is given by:

$$r_{ij} = \left(\sum_{p=1}^{N} (d_{ip} - \mu_i) \cdot (d_{jp} - \mu_j) \right) / \left(\sum_{p=1}^{N} (d_{ip} - \mu_i)^2 \right)^{\frac{1}{2}} \cdot \left(\sum_{p=1}^{N} (d_{jp} - \mu_j)^2 \right)^{\frac{1}{2}}$$

where μ_i and μ_j are the means of X_i and X_j.

<u>step 2:</u> Construct transforming matrix $A = (a_{ij})_{nn}$.

(i) Extract n eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ of S and n corresponding eigenvectors $e_i' = (b_{i1}, \ldots, b_{in})$ for $i = 1, \ldots, n$ such that $\lambda_1 \geq \ldots \geq \lambda_n$.

(ii) Re-scale these n eigenvectors:

$a_{ij} = b_{ij} / (b_{i1}^2 + \ldots + b_{in}^2)^{\frac{1}{2}}$ for $i, j = 1, \ldots, n$.

step 3: Form the new document matrix.

(i) Standardize the document matrix M with respect to each term:

$$d_{ij}'' = (d_{ij} - \mu_i)/\sigma_i$$

where μ_i and σ_i are the mean and the standard deviation of term X_i.

(ii) Premultiply the standardized matrix by the transforming matrix A:

$$M'_{nN} = A_{nn} \cdot M''_{nN}$$

where $M''=(d_{ij}'')$ is the document matrix after standardization.

It should be noted that the eigenvalues are arranged in descending crder in step (2.i). The reason for this arrangement will be discussed in the next section. The re-scaling in step (2.ii) is performed to ensure that the eigenvectors are unique.

Queries can be transformed in the same way as documents. Consider a binary query $Q = (q_1,\ldots,q_n)$ as described in Chapter 2. Q is first transformed to $Q'' = (q_1'',\ldots,q_n'')$ where $q_i''= (q_i-\mu_i)/\sigma_i$, where μ_i, σ_i are the mean and the standard deviation of term X_i. The transformed query $Q'=(q_1',\ldots,q_n')$ is then obtained by multiplying Q" by the transforming matrix A.

3.1.3 The Characterization of the Transformation.

The new terms obtained, besides being uncorrelated, also feature the following properties.

3.1.3.1 The Number of Terms after Transformation.

It is clear that the above transformation will produce the same number of new terms as there are old ones, provided S_{nn} is a nonsingular matrix. If S_{nn} is a singular matrix, a matrix S_{mm} of rank m can be obtained by performing elementary row and column operations on S_{nn}. Hence, m eigenvalues associated with m independent eigenvectors are extracted from S_{mm} and the number of new transformed terms is reduced to m.

Since each component of the eigenvectors is a real number after re-scaling of an eigenvector, every entry of transforming matrix A will be a real number. Furthermore, the document matrix M also becomes a real matrix M" after the standardization of each term. Hence, the above transformation will transfom a binary document matrix to a real matrix M'. We shall treat the real values in the matrix as the weights (importance) of the terms in the documents. Since the terms after the transformation take on real values, the storage is significantly increased in relation to the original binary representation of documents and queries. In addition, the amount of computations required to produce both the actual and expected distributions (to be described in section 3.2) are substantially higher,

especially when the total number of terms is large. Since the number of terms is usually very large (for example, there are 1345 distinct terms in the documents of the CRN4NUL collection), it is often desirable to discard some of newly formed terms.

Intuitively, the terms with small variance are less important than those with large variances. For example, if every document has a particular term with the same importance (weight), this term may be ignored, since this term does not differentiate one document from another. It may be argued that terms with larger variance are more capable of differentiating documents from one another. Hence, if it is required to discard some transformed terms, the terms with smaller variance should be discarded first. By proposition 3.1, the variance of the ith transformed term is equal to the ith eigenvalue of S. Since in the step (2.i) the eigenvalues of S were arranged in descending order, the transformed terms are in descending order according to their variances. Terms with small variance can, therefore, be discarded easily.

3.1.3.2 Other Remarks on the Transformation.

(a). The pair-wise correlations between the newly formed terms have been shown to be zero. Unfortunately, though the uncorrelation between terms does not imply their independence [5]. However, if the original terms set $\{X_1,\ldots,X_n\}$ has multivariate

normal distribution [5], the new terms obtained after transformation $\{Y_1,\ldots,Y_n\}$ are mutually independent and are normally distributed [8,5].

(b). The covariance matrix can be used, instead of the correlation coefficient matrix, while extracting eigenvalues and constructing the transforming matrix. If the covariance matrix is used, then the standardization of the document matrix in step(3) can be omitted. In general, the transformed variables obtained from the correlation coefficient matrix are generally different from those of the covariance matrix.

(C). The transfomation described above can be applied to both binary and real weighted documents and queries.

3.2 The Distribution of Correlations on Weighted Data

We have shown that the weights of terms in the documents after transformation are no longer binary. They are real numbers instead. In the following sections, the actual and the theoretical distributions of the correlations will be derived for the set of real data.

3.2.1 The Actual Distribution of Correlations.

Consider a document matrix $M'_{nN} = (d_{ij}')$ where the rows represent terms and the columns represent documents and the entry d_{ij}' represents the weight of the ith term Y_i in the jth document D_j. Similarly, let a query Q' be represented by a vector $(q_1', \ldots, q_n')$ where q_i' is the weight of the ith term in the query. The correlation of document D_j and query Q is designated by

$$F(D_j, Q') = \sum_{i=1}^{n} d_{ij}' \cdot q_i'.$$

Since the weights of the terms are real numbers instead of binary integers, it is unlikely that any two documents will have exactly the same correlation with respect to a given query. Once all query-document correlations are calculated, we have a set of real correlation values scattered over some range. Hence, we first group these correlation values into a number of intervals and then determine the number of the documents having correlations falling in each of the intervals. The length of the interval is called the interval width. After all correlation values are grouped into intervals and the frequencies in each interval are computed, each individual correlation value may be ignored. The midpoint will be used to represent the correlations of documents for those have query-document correlations falling in that particular interval. Let a_i, for i=1,...,m (number of intervals) be the midpoint of various intervals. A method to decide the interval to which the correlation x of the document belongs is by

$$i = \left\lfloor \frac{x+\frac{1}{2}w}{w} \right\rfloor,$$

where w is a chosen interval width. The frequency of documents having correlation a is obtained by counting the frequency of documents having correlation in each interval i. Thus, the desired distribution of correlations is obtained.

3.2.2 Theoretical Distribution - A Generalized C-function for Weighted Data.

A theoretical distribution similar to the C-function for binary data will be derived in this section for the situation where the weights of terms in documents and queries are real.

Consider the same document matrix $M'=(d_{ij}')$ and query $Q'=(q_1',\ldots,q_n')$ as described in the last section where all weights are real number and all the terms are assumed to be independent of the other terms. We shall first construct the representative as follows:

The values that the weights of a term can assume in documents are grouped into a number of intervals. Associated with each interval, the midpoint of the interval and the relative frequency of th weights of the term in the interval are obtained. For ease of computation, the weights of all the documents in a given interval are assumed to coincide with the midpoint. Thus, for each term we have a distribution of weights in intervals. The representative of

the collection (cluster) is then defined to be the set of these distributions of weights of terms in documents.

The process of generating the theoretical distribution of correlations with respect to the query is presented in two stages: first stage consists of finding the distribution of correlations for each term in the query as if each term was a query with one term and the next stage involves combining these distributions to get the desired theoretical distribution.

To find out the distribution of correlations for term Y_i, we simply multiply the midpoints of various weight intervals of the weight distribution of terms Y_i by q_i', the weights of term Y_i in the query. The distributions of correlations for each term can be represented by a probability generating function, which is a polynomial.

Without loss of generality, we assume that the distributions of correlations for the various terms have the same number of intervals of correlations and the same midpoints of intervals. Since all the terms are assumed to be independent, the distribtution of correlations with respect to the query can then be obtained as the convolution of these distributions correponding to the terms in the query or, equivalently, as the multiplication of all the generating functions.

3.3 Experimental Results

The above tranformation was performed on the CRN4NUL[3] collection. There are 1345 terms in the documents and 467 terms in the queries in the CRN4NUL collection. Since the cost of transformation increases rapidly as the number of terms increases, only the 467 query terms are transformed. The same number of new terms is obtained. In order to avoid the high cost in computing the distribution of query-document correlations, terms whose variances (eigenvalues) are smaller than one are discarded according to the Kaiser-Guttman criterion [5]. As a result, only 178 new terms are obtained.

The 100 queries in the queries collection have been tested. The expected distribution and the actual distribution are produced with respect to a query as decribed in 3.2. The Chi-square test (at the 5% significance level) is then performed to determine how well these two distributions match. 71 out of 100 queries tested satisfy the Chi-square test.

CHAPTER 4

CONCLUSION

It is a common practice to organize documents into clusters in an information retrieval system. While retrieving the desired documents in such an environment, it is often desirable to predict the distribution of query-document correlations of each cluster.

The C-function is found to be suitable for a small collection of similar documents, since terms are likely to be independent in such a collection. Experimental results demonstrate that, for a small collection of documents (for example, the ADIABTH and ADINUL collection with size = 82), the C-function is a good approximation to the actual distribution. It is generally better than the Poisson and the normal distributions for these small collections. The normal distribution would be a good prediction in these small collections if the p_i's are neither too small nor to large. However, in such a situation, the C-function predicts the actual distribution as well as the norma approaches the normal distribution for such p_i's values. The Poisson distribution is not suitable in this situation. In summary, as a predicting distribution of document-query correlations, the C-function is most appropriate for small collections among these three distributions.

Since a large collection of documents is usually not

homogeneous, the probability that a document has a particular term is small. The distribution of correlations for such a collection is likely to be skewed to the ending part (the part with large correlations), since most of the documents have small correlations with the query. Therefore, a symmetric normal distribution should not be used to predict the distribution of correlations for a large collection of documents. Experimental results demonstrate the normal distribution has very poor accuracy in estimating the distribution of correlations for large collection of documents (size $\geq$ 200.) In contrast, the Poisson distribution is more suitable for such a situation. Experimental results have shown that the Poisson distribution has a much better prediction than the normal distribution and the C-function in large collections.

Experimental results show that the modified C-function that takes into consideration the pair-wise dependence of terms shows better prediction than the unmodified C-function. This fact suggests that the terms in the data collection being tested are not totally independent. In order to obtain even better prediction a more complex model (for example, a model that takes into account the dependencies between more than two terms) is required. Another approach is to modify the data collection such that terms are more or less independent. No method of removing the dependencies between terms is available presently unless the original terms are multi-variate normal distributed. However, it is possible to transform a set of dependent

terms to a set of uncorrelated terms. The expected distribution as applied to the newly transformed data is found to be a much better approximation.

Experimental results indicate that terms are more likely to be dependent in a larger collection than in a smaller collection. Therefore, another alternative to handle a large collection is by classifying the collection into small clusters.

Collection	Theoret. dist. % of good fits	C-func.	C-func. with dep. term pairs being combined	C-func. with a classified collection	Poisson dist.	Normal dist.
ADIABTH ND: 82 NQ: 35	34 97.1%	35 100%	34 97.1%	25 71.4%	31 88.6%
ADINUL ND: 82 NQ: 35	31 88.6%	35 100%	33 94.3%	6 17.1%	32 91.3%
CRN2TH ND: 200 NQ: 42	15 35.7%	28 66.6%	19 45.2%	32 76.2%	6 14.3%
CRN2NUL ND: 200 NQ: 42	14 33.7%	31 76.2%	19 45.2%	33 78.5%	3 7.1%
CRN4NUL ND: 424 NQ: 155	27 17.4%	70 45.2%	41 26.5%	86 55.7%	2 1.3%
MEDNUL ND: 450 NQ: 30	12 40.0%	23 76.7%	19 63.3%	24 80.0%	1/29 3.4%
CRN14TH ND: 1400 NQ: 225	11 4.9%	46 20.4%	16 7.1%	77 34.2%	0 0%

Table 1. Results of the Chi-square test of goodness of fit at the 5 % significance level and (k-1) degrees of freedom (k = no. of cells.)

ND: no. of documents in the collection
NQ: no. of queries (= no. of tests) in the collection
* : no. of good fits
**: one of the tests fails due to degree of freedom < 1

Query no.	No. of terms	No. of term pairs	average Chi-square quantity	No. of dep. term pairs	Average Chi-square quantity of dep. term pairs
23	4	6	6.0	2	15.3
30	6	15	6.0	4	19.6
41	8	28	11.6	12	26.2
47	8	28	35.0	19	51.3
59	4	6	155.8	4	233.1
75	13	78	10.6	19	41.1
80	10	45	3.7	11	11.1
93	11	55	7.0	19	18.1
105	8	28	28.0	15	51.4
115	11	55	3.5	13	11.5
119	12	66	8.1	16	30.8
113	4	6	34.5	2	102.3
139	12	66	18.4	23	50.7
140	4	6	16.5	3	31.8
153	4	6	27.0	2	80.1
160	17	136	15.1	49	40.5
176	8	28	10.0	9	29.3
185	4	6	33.7	4	50.0
210	6	15	60.0	6	148.0
224	10	45	11.0	17	27.5

Table 2. Sample results (from the CRN14TH collection) of 2 by 2 contingency table test for independence of term pairs of queries at the 5% significance level and 1 degree of freedom.

The critical value is 3.8.

Sample size				
% of good fits	82	200	400	1400
Theoret. dit.				
The C-function	161 71.6%	115 51.1%	71 31.6%	11 4.9%

Table 3. Chi-square test of goodness of fit on samples of the CRN14TH collection.

No. of queries (= no. of tests) = 225.

Bibliography

[1] Salton, G. Dynamic information and library processing. Prentice Hall, Inc. N.J., 1975.

[2] Yu, C.T. and Luk, W.S. Analysis of effectiveness of retrieval in clustered files. J.ACM 24, 4 (Oct. 1977), 607-622.

[3] SALTON, G. The SMART retrieval system - experiments in automatic document processing. Prentice Hall, Englewood Cliffs, N.J., 1971.

[4] Feller, W. An introduction to probability theory and its application. Vol. 1 & 2, 2nd edition, John Wiley & Sons, Inc. 1957.

[5] Afifi, A.A. and Azen, S.P. Statistical analysis - A computer oriented approach. Academic Press, Inc. N.Y. 1974.

[6] Dorf, R.C. Matrix algebra - A programmed introduction. John Wiley & Son, Inc. 1969.

[7] Tropper, A.M. Matrix theory for electrical engineering students, George G. Harrap & Co. LTD. london, 1962.

[8] Lancaster, H.O. Zero correlation and independence.

Australian Journal of Statistics. Vol. 1, 1959.

[9] Damerau, J. An experiment in automatic indexing. Amer. Document. 16,4(Oct. 1965), 283-289.

[10] Harter, S. A probabilistic approach to modern keyword indexing, Part 1. J. Amer. Soc. Inform. Sci. 26, 4(July-Aug. 1975), 197-206.

[11] Swets, J. Information retrieval systems, Science 141 (1963), 245-250.

[12] Borko, B. and Bernick, M. Automatic document classification. J. ACM 10 (1962), 151-162.

[13] Borko, B. and Bernick, M. Automatic document classification, Part II. Additional Experiments. J .ACM 11(1963), 138-151.

[14] Bookstein, A. and Kraft D. Operations research applied to document indexing and retrieval decisions. J.ACM 24(1977), 418-427.

[15] Cochran, W. Chi-square test of goodness of fit. Annals of American Mathematical Statistics. 23(1952),315-345.

[16] Berkson, J. Some difficulties of interpretation encountered in application of Chi-square test. J.

Amer. Stat. Assn. 33(1938), 526-536.

[17] Fisz M. Probability theory and mathematical statistics. 3rd edition, John Wiley & Son, Inc. N.Y.,1967.

[18] Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos. Mag. Series 5, vol 50(1900) 157-172.

[19] Bookstein, A. and Swanson D.R. Probabilistic models for automatic indexing. J. Amer. Soc. Inform. Sci. 25, 4(1974), 312-318.